Wharton

Financial Institutions Center

Analyzing Firm Performance in the Insurance Industry Using Frontier Efficiency Methods

by
J. David Cummins
Mary A. Weiss

98-22

The Wharton School
University of Pennsylvania

THE WHARTON FINANCIAL INSTITUTIONS CENTER

The Wharton Financial Institutions Center provides a multi-disciplinary research approach to the problems and opportunities facing the financial services industry in its search for competitive excellence. The Center's research focuses on the issues related to managing risk at the firm level as well as ways to improve productivity and performance.

The Center fosters the development of a community of faculty, visiting scholars and Ph.D. candidates whose research interests complement and support the mission of the Center. The Center works closely with industry executives and practitioners to ensure that its research is informed by the operating realities and competitive demands facing industry participants as they pursue competitive excellence.

Copies of the working papers summarized here are available from the Center. If you would like to learn more about the Center or become a member of our research community, please let us know of your interest.

Anthony M. Santomero
Director

The Working Paper Series is made possible by a generous grant from the Alfred P. Sloan Foundation

Analyzing Firm Performance in the Insurance Industry Using Frontier Efficiency Methods

By

J. David Cummins and Mary A. Weiss

July 1998

J. David Cummins
The Wharton School
3641 Locust Walk
Philadelphia, PA 19104-6218
Phone: 215-898-5644
Fax: 215-898-0310
Email: **cummins@wharton.upenn.edu**

Mary A. Weiss
Department of Risk Management
and Insurance
Temple University
Ritter Hall Annex
Phone: 215-204-1916
Fax: 610-520-9069
Email: mweiss@vm.temple.edu

Analyzing Firm Performance in the Insurance Industry
Using Frontier Efficiency Methods

1. Introduction

In analyzing insurance firms, it is often important to measure their performance relative to other firms in the industry. Traditionally, this has been done using conventional financial ratios such as the return on equity, return on assets, expense to premium ratios, etc. With the rapid evolution of frontier efficiency methodologies, the conventional methods are rapidly becoming obsolete. Frontier methodologies measure firm performance relative to "best practice" frontiers consisting of other firms in the industry. In the future, tests of economic hypotheses about insurers about such matters as organizational form, distribution systems, economies of scale and scope, and the effects of mergers and acquisitions will not be convincing unless they involve the use of one or more frontier-based performance measures. Such measures dominate traditional techniques in terms of developing meaningful and reliable measures of firm performance. They summarize firm performance in a single statistic (for a given type of efficiency) that controls for differences among firms in a sophisticated multidimensional framework that has its roots in economic theory.

Most efficiency analyses to date in insurance and elsewhere have focused on production and cost efficiency. More recently, researchers have begun to estimate revenue and profit frontiers. Perhaps the most basic frontier is the production frontier, which is estimated based on the assumption that the firm is minimizing input use conditional on output levels.[1] Production frontiers can be estimated even if data on input and output prices are unavailable. If data on input prices are available, it is possible to estimate the cost frontier, usually based on the assumption that the firm is minimizing costs conditional on output levels and input prices. Ultimately, of course, the firm also can optimize by choosing its level of output and/or output mix. The revenue and profit functions allow the firm to do this by maximizing revenues or

[1]This definition applies to an *input-oriented* frontier. It is also possible to develop output-oriented measures of efficiency by maximizing outputs conditional on inputs. Most efficiency analysis to date in insurance and other financial services industries have been input-oriented, and most of our discussion in the paper assumes an input-orientation.

profits, respectively, contingent only on input and output prices.[2] Finally, sophisticated methods such as Malmquist analysis have recently been developed for measuring changes in efficiency and shifts in the frontier over time.[3]

Frontier efficiency methods are useful in a variety of contexts. One important use is for testing economic hypotheses. For example, both agency theory and transactions cost economics generate predictions about the likely success of firms with different characteristics in attaining objectives such as cost minimization or profit maximization under various economic conditions. Firm characteristics that are likely to be important include organizational form, distribution systems, corporate governance, and vertical integration. Frontier methodologies have been used to analyze a wide range of such hypotheses.[4]

A second important application of frontier methodologies is to provide guidance to regulators and policy makers regarding the appropriate response to problems and developments in an industry or the economy in general. For example, both the banking and insurance industries are currently experiencing a wave of mergers and acquisitions. Frontier methodologies can be used to determine whether consolidation is likely to be beneficial or detrimental in terms of the price and quality of services provided to consumers. From time to time, the efficiency of insurer operations also becomes an important regulatory issue, as in the debate over the price of automobile insurance. Frontier efficiency methods also can be used to shed light on this type of problem.

---

[2]Frontier analysis is typically conducted under the assumption that the industry is competitive. However, it is also possible to test the hypothesis of competitiveness and to measure efficiency for non-competitive industries. One departure from the usual competitive assumptions is provided by a public entity such as a public utility or government agency. Such institutions have been widely studied using frontier methodologies. For examples, several chapters in Charnes, et al. (1996) and Fried, Lovell, and Schmidt (1993) deal with public entities.

[3]Although the terms *productivity* and *efficiency* are often used interchangeably, they are in fact different economic concepts. Simply put, efficiency refers to how well firms are performing relative to the existing technology in an industry; whereas productivity refers to the evolution of technology over time. Frontier efficiency methods are available for measuring both efficiency and productivity.

[4]Berger and Humphrey (1997) provide a review of applications to financial institutions.

A third application of frontier methodologies is to compare economic performance across countries. For example, Fare, et al. (1994) compare the evolution of productivity in industrialized nations. Weiss (1991b) compares productivity in the property-liability insurance industries of the U.S. and four European countries, and recent studies have also compared banking efficiency in the U.S. and in various European nations (e.g., Pastor, Pérez, and Quesada, 1997).

A fourth application is to inform management about the effects of policies, procedures, strategies, and technologies adopted by the firm. Although firms currently employ a variety of benchmarking techniques, frontier analysis can provide more meaningful information than conventional ratio and survey analysis, which often overwhelms the manager with masses of statistics that are difficult to summarize conveniently in terms of one or a few performance measures. Frontier analysis can be used not only to track the evolution of a firm's productivity and efficiency over time but also to compare the performance of departments, divisions, or branches within the firm.

The purpose of this chapter is to provide an overview of frontier efficiency methodologies, a discussion of methodological issues specific to insurance, and a review of applications to the insurance industry. Section 2 provides an overview of the principal frontier methodologies. Section 3 discusses the measurement of inputs and outputs as well as some additional methodological issues and problems. Section 4 provides a review of the efficiency literature in insurance, and section 6 concludes.

2. Frontier Methodologies for Estimating Efficiency and Productivity

There are two principal types of efficiency methodologies – the *econometric (parametric) approach* and the *mathematical programming (non-parametric)* approach. The econometric approach requires the specification of a production, cost, revenue, or profit function as well as assumptions about the error term(s). The primary advantage of the econometric approach is that it allows firms to be off the due to random error as well as inefficiency. However, this methodology is vulnerable to errors in the specification of the functional form or error term(s). The mathematical programming approach avoids this type of specification error by imposing somewhat less structure on the optimization problem, i.e., neither functional form nor error term

assumptions are required. However, in most applications of the methodology, any departure from the frontier is measured as inefficiency, i.e., random error or bad luck is not separated out. Both approaches have advocates and neither has emerged as dominant. Some recent papers recommend using more than one methodology to check the robustness of the results.[5] As methodologies continue to evolve, it is likely that the major limitations of the methods will be overcome. For example, recent papers have begun to develop stochastic mathematical programming models (Land, Lovell, and Thore 1993) as well as providing the underlying statistical theory for the mathematical programming (e.g., Grosskopf, 1996). This section introduces the concept of economic efficiency and then discusses the two types of estimation methodologies.

The Concept of Economic Efficiency

The concept of economic efficiency flows directly from the microeconomic theory of the firm. Perhaps the most basic concept is that of the *production frontier*, which indicates the minimum inputs required to produce any given level of output for a firm operating with full efficiency. A production frontier for a firm with one input and one output is shown in Figure 1. If firm i is operating at point (x_i, y_i), it could operate more efficiently by moving to the frontier, i.e., by adopting the state-of-the-art technology. The firm's level of *technical efficiency* is given by the ratio 0a/0b, which is the reciprocal of its distance from the frontier, 0b/0a.

If the firm has more than one input, inefficiency can also result from the firm's not using the cost minimizing combination of inputs. This type of inefficiency, known as *allocative inefficiency***,** is shown in Figure 2, which illustrates Farrell's (1957) technical and allocative efficiency concepts. The diagram shows an isoquant for a firm with one output and two inputs, x_1 and x_2. The isoquant QQ' in Figure 2 represents the various combinations of the two inputs required to produce a fixed amount of the single output using the best available technology. Thus, firms operating on the isoquant are considered to be technically efficient. The

[5]Cummins and Zi (1998) apply a variety of econometric and non-parametric techniques to estimate cost efficiency for a sample of U.S. life insurers. They find that econometric efficiency estimates are robust to the choice of distributional assumptions from the error term. The rank correlations among efficiency scores for the econometric methods are typcially above 0.95. The rank correlations between the econometric and mathematical programming efficiency estimates are lower (around 0.67), but the results of the two approaches are generally consistent. For a similar analysis of the banking industry see Bauer, et al. (1998).

optimal operating point is represented by the tangency (point D) between the isoquant QQ′ and the isocost line ww′. A firm operating at this point is considered to be fully *cost efficient*. The firm operating at point A = ($x_1^A x_2^A$) exhibits both technical and allocative inefficiency. It is technically inefficient because it is not operating on the best-technology isoquant. The measure of Farrell technical efficiency is the ratio 0B/0A, i.e., the proportion by which it could radially reduce its input usage by adopting the best technology. However, this firm is also allocatively inefficient because it is not using its inputs in the correct proportions. Specifically, it is using too much of input 2 and not enough of input 1. The measure of allocative efficiency is thus the ratio 0C/0B. Cost efficiency is then defined as follows:

Cost Inefficiency = Technical Efficiency * Allocative Efficiency = (0B/0A) * (0C/0B) = 0C/0A

It is also possible to decompose technical efficiency into two components: *pure technical efficiency* and *scale efficiency*. These concepts are illustrated in Figure 3, which shows two production frontiers for the single input-single output case. Frontier V^C represents a constant returns to scale (CRS) frontier, while frontier V^V is a variable returns to scale (VRS) frontier. It is socially and economically optimal for firms to operate at constant returns to scale, providing the motivation for separating pure technical and scale efficiency. Consider firm i, operating at point (x_i, y_i). Pure technical efficiency is measured relative to the VRS frontier and is equal to 0b/0c. This is the proportion by which the firm could reduce its input usage by adopting the best technology. However, a firm operating on the VRS frontier at firm i's output level is also scale inefficient because it is not operating on the CRS frontier. Its scale efficiency is measured by the ratio 0a/0b. Thus, we can define:

Technical Efficiency = Pure Technical Efficiency * Scale Efficiency

= (0b/0c)*(0a/0b) = 0a/0c

The production frontier also can be used to illustrate productivity progress or regress. For example, consider Figure 4, which shows production frontiers for periods t and t+1 (V^t and V^{t+1}, respectively) for the single input-single output firm. The frontier for period t+1 lies to the left of the frontier for period t. This implies that productivity gains have been achieved between periods t and t+1. Consider a specific firm operating at point (x_i^t, y_i^t) in period t and at point (x_i^{t+1}, y_i^{t+1}) in period t+1. This firm has become both more

productive and more efficient between periods t and t+1. In period t+1, the firm is operating at a level of output that would have been infeasible in period t, i.e., its operating point lies to the left of the production frontier for period t. In addition, the firm's efficiency has improved because it is closer to the frontier in period t+1 than it was in period t. The type of inefficiency considered here is *technical efficiency*, i.e., firms on the frontier are using the most efficient available technology, while those to the right of the frontier are not using this technology.

This discussion can be formalized by reference to the input distance function introduced by Shephard (1970). Suppose producers use input vector $x = (x_1, x_2,... x_k) \in \mathbb{R}_+^k$ to produce output vector $y = (y_1, y_2, ..y_n) \in \mathbb{R}_+^n$. A production technology which transforms inputs into outputs can be modeled by an input correspondence $y \rightarrow V(y) \subseteq \mathbb{R}_+^k$. For any $y \in \mathbb{R}_+^n$, V(y) denotes the subset of *all* input vectors $x \in \mathbb{R}_+^k$ which yield at least y. V(y) is assumed to satisfy certain axioms (see Färe, Grosskopf, and Lovell, 1985, and Färe, 1988) . The input oriented distance function is defined by

$$D(x,y) = sup\{ \theta : (\frac{x}{\theta}, y) \in V(y) \} = (inf\{ \theta : (\theta x, y) \in V(y) \})^{-1} \qquad \textbf{(1)}$$

The input distance function is the same as the reciprocal of the minimum equi-proportional contraction of the input vector x, given outputs y, i.e., Farrell's (1957) measure of input technical efficiency. Input technical efficiency TE(x,y) is therefore defined as TE(x,y) = 1/D(x,y). TE(x,y) for each decision making unit can be obtained by linear programming (Charnes, et al., 1994).

To illustrate the distance function, consider the firm operating at point (x_i^t, y_i^t) in Figure 4. The distance function value for this firm is given by $D^t(x_i^t, y_i^t)$ = 0a/0b, where superscripts on D indicate the time period of the frontier from which the distance is computed. Distance functions can be used to compare the firm's efficiencies in periods t and t+1. In Figure 4, $D^{t+1}(x_i^{t+1},y_i^{t+1})$ = 0e/0f < $D^t(x_i^t,y_i^t)$ = 0a/0b. The distance function representation also can be used to define the *Malmquist index* of total factor productivity. If our interest is in determining whether productivity change has occurred between periods t and t+1, we could use

either the period t frontier or the period t+1 frontier as our point of reference. With respect to the period t frontier, an input-oriented Malmquist productivity index can be defined as:

$$M^t = \frac{D^t(x^t, y^t)}{D^t(x^{t+1}, y^{t+1})} \quad \textbf{(2)}$$

The input-oriented Malmquist index with respect to the period t+1 frontier is:

$$M^{t+1} = \frac{D^{t+1}(x^t, y^t)}{D^{t+1}(x^{t+1}, y^{t+1})} \quad \textbf{(3)}$$

To avoid arbitrarily choosing one frontier to compute the index, the usual approach is to take the geometric mean, yielding the following *Malmquist index of total factor productivity* (Grosskopf, 1993):

$$M(x^{t+1}, y^{t+1}, x^t, y^t) = \left[\frac{D^t(x^t, y^t)}{D^t(x^{t+1}, y^{t+1})} \frac{D^{t+1}(x^t, y^t)}{D^{t+1}(x^{t+1}, y^{t+1})} \right]^{1/2} \quad \textbf{(4)}$$

This expression can be factored into two components, representing *efficiency change*, i.e., the change in Farrell technical efficiency between the two periods, and *technical change*, i.e., the shift in the frontier between the two periods. The decomposition is illustrated in Figure 4. Efficiency change is the ratio of the distance from the frontier in period t to the distance in period t+1, i.e., $D^t(x_i^t, y_i^t)/D^{t+1}(x_i^{t+1}, y_i^{t+1}) = [(0a/0b)/(0e/0f)]$. If technical efficiency has improved between year t and year t+1, the ratio will be greater than 1. Technical change is measured by comparing the input-output bundle in period t+1 to both the period t+1 and period t technologies, and likewise for the input-output bundle in year t. Technical change is then computed as follows:

$$\textit{Technical Change} = \left[\frac{D^{t+1}(x^{t+1}, y^{t+1})}{D^t(x^{t+1}, y^{t+1})} \frac{D^{t+1}(x^t, y^t)}{D^t(x^t, y^t)} \right]^{1/2} = \left[\left(\frac{0e/0f}{0e/0d} \right) \left(\frac{0a/0c}{0a/0b} \right) \right]^{1/2} \quad \textbf{(5)}$$

Intuitively, if favorable technical change has occurred, the frontier will have moved to the left, and both output bundles will be further from the period t+1 frontier than they are from the period t frontier. Thus, a ratio greater than 1 indicates favorable technical change. The product of technical efficiency change and technical change is total factor productivity change.

Econometric Frontier Efficiency Models

The two most important decisions that must be made in applying the econometric frontier efficiency methodology are the choice of functional form and the treatment of the error term. This section first discusses the functional forms that are used most frequently and then turns to a discussion of specifying and estimating the error term.

Functional Form. Ideally, researchers would be able to determine the exact form of the production function for the firms being analyzed. This is, in fact, possible for some physical production processes such as manufacturing chemicals or refining oil. However, in most industries, and especially in the service sector, the exact functional form is not known. In the past, this led economists to use various approximations such as the well-known Cobb-Douglas and constant elasticity of substitution (CES) production functions. One of the most important developments in the evolution of parametric frontier modeling was the introduction of the translog production function by Christensen, Jorgenson, and Lau (1973). They reasoned that even though the functional form may be unknown, any function satisfying rather weak regularity conditions can be expanded as a single or multi-variate Taylor series. They proposed the use of a second-order Taylor expansion in natural logarithms as an approximation of the unknown production function. A directly analogous derivation leads to the translog cost function. The translog has an advantage over earlier functional forms in that it allows returns to scale to change with output or input proportions so that the estimated cost curve can take on the familiar U-shape. The quadratic feature of the translog is also a potential disadvantage, as explained below.

A general expression of a cost function is $C = f(Y,W,t)$, where C is total cost, Y is output, W is input price, and t is time. In most applications, Y and W are vectors. The *cost frontier* is defined as the minimum total cost function, i.e., the function that gives the minimum attainable cost for each level of output. The cost frontier is denoted $C^F = C^F(Y,W,t)$. The translog cost function is:

$$\ln C_{st} = [\alpha_0 + \sum_{i=1}^{n} \alpha_{Yi} \ln Y_{sit} + \frac{1}{2} \sum_{i=1}^{n} \sum_{k=1}^{n} \alpha_{Yik} \ln Y_{sit} \ln Y_{skt} + \sum_{j=1}^{m} \alpha_{wj} \ln w_{sjt} + \frac{1}{2} \sum_{j=1}^{m} \sum_{f=1}^{m} \alpha_{wjf} \ln w_{sjt} \ln w_{sft} + \sum_{i=1}^{n} \sum_{j=1}^{m} \alpha_{yiwj} \ln Y_{sit} \ln w_{sjt}] + v_{st} + \epsilon_{st} \quad \textbf{(6)}$$

where s = {1, ..., S}, i = {1, ..., n}, and j = {1, ..., m} index firms, outputs, and inputs, respectively, C_{st} = observed total costs for firm s in year t = $\Sigma_j w_{sjt} X_{sjt}$, X_{sjt} = quantity of input j used by firm s in year t, Y_{sit} = amount of output i produced by firm s in year t, w_{sjt} = price of input j to firm s in year t, ϵ_{st} = a random error term, and v_{st} = an inefficiency error term. The estimation is usually conducted as a system of equations consisting of the cost function and the first-order conditions for cost minimization:

$$\frac{\partial \ln C_{st}}{\partial \ln w_{sjt}} = \frac{w_{sjt} X_{sjt}}{C_{st}} = [\alpha_{wj} + \sum_{f=1}^{m} \alpha_{wjf} \ln w_{sft} + \sum_{i=1}^{n} \alpha_{yiwj} \ln Y_{sit}] + \omega_{sjt} \quad \textbf{(7)}$$

where ω_{sjt} = a random error term. Linear homogeneity and symmetry restrictions are imposed in the estimation.

Firms are assumed to share a common cost frontier given by the bracketed expression in equation (6). The stochastic nature of the frontier is modeled by adding a two-sided random error term, ϵ_{st}, to the cost equation. The realizations of these random errors differ across firms, but the errors are assumed to be independent, identically distributed, and beyond the control of individual firms. Hence, ϵ_{st} is not indicative of inefficiency. Inefficiency is captured by the additional error term in equation (6), v_{st} . Because inefficiency can only increase (not reduce) costs, v_{st} is a one-sided error term, $v_{st} \geq 0$, or more generally $v_{st} \geq \zeta$, where ζ = a non-negative parameter. The input shares are assumed to have a functional component which is common to all firms (the bracketed expression in equation (7)) and a random component captured by the two-sided error term ω_{sjt}, where $\Sigma_j \omega_{sjt} = 0$.

While the translog has been widely used in econometric efficiency studies, it has some limitations that have led some researchers to use alternative forms for the cost function. One limitation is that the translog does not naturally allow any of the independent variables to be equal to zero. Although this is not a problem with regard to input prices, it can be a limitation for outputs if more than one output is present and some firms

do not produce all outputs. This is especially problematical in studying economies of scope, where zeros for some outputs are required to obtain meaningful results.

When zero outputs are present, one approach is to salvage the translog using somewhat ad hoc techniques such as setting all zero outputs to a small positive number or adding 1 to the value of all outputs (not just the output involving the zeros). The approach of setting zero outputs to a small positive number has been shown to be unsatisfactory in studies of scope economies because quite different estimates of scope economies can be obtained, depending upon how close the number is to zero (e.g., Röller, 1990).

Although the ad hoc techniques have acquired a respectable history outside of the scope literature, for many purposes it may be advisable to use an alternative functional form. We discuss three alternatives that show up relatively often in the financial services literature. The simplest is the Fuss normalized quadratic, which replaces the logged values of outputs and input prices in equation (6) with the unlogged values of the variables (Morrison and Berndt, 1981) . Homogeneity is imposed by dividing all variables by one of the input prices. A limitation of this form is that the results are not completely invariant to which input is chosen for normalization. An alternative is the generalized translog cost function, obtained by transforming the output variables using a Box-Cox transformation (Caves, Christensen, and Tretheway, 1980). I.e., the $\ln(Y_{it})$ in equation (6) are replaced by the Box-Cox transformed variate defined as $Y_{it}^{(\phi)} = (Y_{it}^{\phi} - 1)/\phi$, $\phi \neq 0$. The Box-Cox model is the same as the translog if $\phi = 0$; and, as a result, the approach fails to improve on the translog if ϕ is close to zero.

Another functional form that seems ideally suited to the analysis of scope economies is the *composite cost function* (Pulley and Braunstein, 1992, Pulley and Humphrey, 1993). This functional form consists of a quadratic component for outputs, linked through interaction terms with a log-quadratic component for input prices. The resulting functional form can be estimated linearly, log-linearly or using a Box-Cox transformation. This functional form has been used by Berger, Cummins, and Weiss (1998) to analyze economies of scope in the insurance industry, considering firms that specialize in either life or property-liability insurance along with those that write both types of insurance.

A limitation of all of the quadratic cost functions, including the translog and composite functions, is that they force the cost function to be U-shaped. This may be a problem if, for example, the actual cost curve exhibits constant returns to scale after output reaches the level where firms are no longer operating in the range of increasing returns to scale. The problem arises because the translog was developed as a local approximation to the true underlying cost function and thus may give misleading results when used globally. This problem cannot be solved by extending the Taylor series expansion to include cubic or higher terms because the resulting function is still a local approximation. Several approaches have been proposed for solving this problem (see McAllister and McManus, 1993). A particularly promising approach is the use of the Fourier flexible functional form, first proposed by Gallant (1982).[6] This form arises from the expansion of the unknown true cost function as a Fourier series. The usual procedure is to append the Fourier (sine and cosine) terms to a standard translog, giving an extremely flexible function that will not force the estimated cost function to have a region characterized by decreasing returns to scale.

Mitchell and Onvural (1992) and McAllister and McManus (1993) find the Fourier form to be superior to the translog in estimating bank cost functions. In the only insurance application of the function to date, Berger, Cummins, and Weiss (1997) find the Fourier terms in their cost and profit function models to add significant explanatory power to the translog. However, the results of their hypothesis tests were the same under the translog and the Fourier functional forms.

Separating Inefficiency and Random Error. The usual distributional assumptions are normal distributions for ϵ_{st} and ω_{sjt} (see equations (6) and (7)) and a truncated normal, exponential, or gamma distribution for v_{st}.[7] The general procedure for estimating efficiency using equation (6) is to estimate the cost function parameters and $z_i = \epsilon_i + v_i$ by maximum likelihood and then to calculate efficiency for each observation

[6]See Mitchell and Onvural (1992) for an application to banking and Berger, Cummins, and Weiss (1997) for an application to insurance.

[7]For specificity, this discussion focuses on the translog, but a similar approach would apply for the other functional forms discussed above.

in the sample by separating the inefficiency component from the total estimated error term. This involves finding the conditional probability distribution of v_i given z_i and finding the conditional expectation $E(\exp(-v_i)|z_i)$ (see Greene, 1993)), providing an estimate of the ratio of frontier costs to actual costs for each firm in the sample.

An alternative to making explicit distributional assumptions is provided by the *distribution free* method developed by Schmidt and Sickles (1984) and Berger (1993). This method can be used when several years of data are available. The cost function is estimated for the entire data period, either year by year or by pooling the data for all years. The residuals from the cost function estimation constitute a vector of random error terms for each firm, $\mathbf{z_i} = \{z_{i1}, z_{i2}, \ldots, z_{iT}\}$, i = 1, 2, . . . , S. The error term z_{it} is specified here as $z_{it} = \epsilon_{it} + v_i$, i.e., the inefficiency component is assumed to be the same for all years. No distributional assumptions are imposed on ϵ_{it} or v_i. Rather, an estimate of the efficiency is extracted by averaging the estimated overall error, $z_{it} = v_i + \epsilon_{it}$, over the sample period on the assumption that the random error ϵ_{it} will average out over time. Cost efficiency is then estimated for each firm as:

$$E[v_i \mid z_{11} \ldots z_{TS}] = \exp(\min_i(\bar{z}_i) - \bar{z}_i) \qquad \textbf{(8)}$$

where $\bar{z}_i$ denotes the average over the sample period of the residuals z_{it} for firm *i*, and $\min_i(\bar{z}_i)$ is the minimum average error term for the firms in the sample. In addition to avoiding distributional assumptions, this method is also easier to implement than the distributional approach because it does not require the use of maximum likelihood methods. A problem with the method is that it may give misleading results if the inefficiency component of the error term is not constant over time or if the number of available data years is not sufficient to average out the random error.

Mathematical Programming Methods

The mathematical programming (non-parametric) approaches to estimating efficiency represent an empirical implementation of Shepard's distance function methodology discussed above. The implementation that is used most frequently is *data envelopment analysis (DEA)*. The method can be used to estimate production, cost, and revenue frontiers and provides a particularly convenient way for decomposing efficiency

into its components. E.g., cost efficiency can be conveniently decomposed into pure technical, scale, and allocative efficiency. Intuitively, the method involves searching for a convex combination of firms in the industry that dominate a given firm. These firms constitute the given firm's *reference set.* If the reference set consists only of the firm itself, it is considered self-efficient and has an efficiency score of 1.0. However, if a dominating set can be found consisting of other firms, the firm's efficiency is less than 1.0. The implication is that the firm's outputs could be produced more cheaply (in the case of cost efficiency) by the "best practice" firms in the industry.

DEA efficiency is estimated by solving linear programming problems. For example, technical efficiency is estimated by solving the following problem, for each firm, i = 1,2,... S, in each year of the sample period (time superscripts are suppressed):

$$(D(y_i, x_i))^{-1} = T(y_i, x_i) \qquad \textbf{(9)}$$

$$= \min \theta_i$$

$$\textit{subject to}: \quad Y\lambda_i \geq y_i$$

$$X\lambda_i \leq \theta_i x_i$$

$$\lambda_i \geq 0$$

where **Y** is an N x S output matrix and **X** a M x S input matrix for all firms in the sample, $\mathbf{y}_i$ is a N x 1 output vector and x_i an M x 1 input vector for firm i, and λ_i is an S x 1 intensity vector (the inequalities are interpreted as applying to each row of the relevant matrix). The constraint $\lambda_i \geq 0$ imposes constant returns to scale. The firms for which the elements of λ_i are non-zero constitute the firm i's reference set.

Technical efficiency is separated into pure technical and scale efficiency by reestimating problem (9) with the additional constraint

$$\sum_{i=1}^{I} \lambda_i = 1$$

for a variable returns to scale (VRS) frontier (this step estimates pure technical efficiency), and again with the

constraint

$$\sum_{i=1}^{I} \lambda_i \leq 1$$

for a non-increasing returns to scale (NIRS) frontier. Pure technical efficiency is defined as the distance from the variable returns to scale frontier (see Figure 3), and the relationship $TE(x_i,y_i) = PT(x_i,y_i)\ S(x_i,y_i)$ can be used to separate pure technical and scale efficiency, where $S(x_i,y_i)$ represents scale efficiency and $PT(x_i,y_i)$ pure technical efficiency. Thus, if TE = PT, i.e., the CRS and VRS technical efficiency estimates are equal, then $S(x_i,y_i) = 1$ and CRS is indicated. If $S \neq 1$ and the NIRS efficiency measure = PT, DRS is present; whereas if $S \neq 1$ and the NIRS efficiency measure $\neq$ PT, then IRS is indicated (Aly, et al., 1990).

A two-step procedure is used to estimate DEA cost efficiency. The first step is to solve the following problem:

$$\underset{x_i}{Min} \quad w_i^T x_i \qquad \textbf{(10)}$$

subject to: $Y\ \lambda_i \geq y_i$, k = 1,2,...,K,

$X\ \lambda_i \leq x_i$, n = 1,2,..,N,

$\lambda_i \geq 0$, j = 1,2,.. S,

where T stands for vector transpose. The solution vector x_i^* is the cost-minimizing input vector for the input price vector w_i and the output vector y_i. The second step is to calculate firm i's cost efficiency as the ratio $\eta_i = w_i^T x_i^* / w_i^T x_i$, i.e., the ratio of frontier costs to actual costs. Thus, cost efficiency satisfies the inequality, $0 < \eta_i \leq 1$, with a score of 1 indicating that the firm is fully cost efficient.

Revenue efficiency is estimated similarly to cost efficiency. However, in this case we adopt an output-oriented rather than an input-oriented approach and maximize revenues rather than minimizing costs. The setup of the problem is suggested by Lovell (1993). Specifically, the following problem is solved for each firm in each year of the sample period:

$$\underset{Y_i}{Max} \sum_{n=1}^{N} p_{ni} y_{ni} \qquad \textbf{(11)}$$

Subject to $x_{ki} \geq \sum_j \lambda_j x_{kj}$ k = 1,2,...,K,

$y_{ni} \leq \sum_j \lambda_j y_{nj}$, n = 1,2,..,N, and

$\lambda_j \geq 0$, j = 1,2,.. S.

The solution vector Y_i^* is the revenue maximizing output vector for the output price vector p_i and the input vector X_i. Revenue efficiency is then measured by the ratio $\kappa_i = p_i^T Y_i / p_i^T Y_i^* \leq 1$. Linear programming is used to solve the problem defined in (11).

All of the DEA methods discussed so far impose the condition that the efficient frontier be a convex set. While this generally seems to be a reasonable assumption, there is no necessary mathematical or economic reason why it should always hold in practice. Deprins, Simar, and Tulkens (1984) criticize the DEA methodology for imposing the convexity assumption, contending that it leads to a poor fit to some observed data sets because it does not allow for local non-convexities. Intuitively, the convexity assumption allows a firm to be dominated by a convex combination of other firms even if there is no firm actually operating with the input-output vector of the "virtual" firm created by the convex combination. Deprins, Simar, and Tulkens (1984) propose the elimination of the convexity assumption, leading to the *free disposal hull (FDH)* estimation technique. The FDH name comes from its retention of another major assumption of DEA, free disposability, which implies, for example, that outputs do not decrease if some inputs are increased (strong disposability of inputs). The FDH method allows the frontier to have local non-convexities. It has been shown to envelop the data more closely than DEA, and FDH efficiencies tend to be considerably higher than those for DEA with many more self-efficient firms (Eeckaut, Tulkens, and Jamar, 1993, Cummins and Zi, 1998). However, it is not at all clear that the increase in goodness of fit is economically desirable, i.e., the frontier may indeed be convex for some industries. More research is clearly needed to resolve the convexity issue.

The Index Approach

The index approach provides an alternative to the econometric and mathematical programming methods. Under the index approach, total factor productivity (TFP) growth is defined as the difference between output and input growth. Also, indexes are used sometimes to condense the number of outputs and/or inputs for a multi-output or multi-input firm. To use this approach, data for output and input quantities and prices are required. No parameters are estimated, but the index formula itself usually is derived from an assumed functional form for cost or production. The most popular index by far is the Divisia index (Diewert, 1981).

The Divisia index can be derived from a translog aggregator (flexible) function exhibiting constant returns to scale and profit maximizing competitive behavior. When used to measure productivity, productivity growth is assumed to be Hicks neutral. The "exact" index may be used if non-constant returns to scale are known to exist (Diewert, 1981). In cases where these assumptions are not reasonable, ex post regression analysis may be used to isolate the effect of such factors as size and regulatory impacts.

The index approach is used typically in cases where direct econometric estimation of a cost or production function is infeasible because the functional form for cost or production is not known and/or a sufficient number of observations to estimate the numerous parameters in flexible functional forms are not available. It is not uncommon for this approach to be used in analyzing national accounting data, such as insurance gross product originating because it is easy to compute (i.e., no estimation is conducted) (e.g., see Bernstein 1997).

3. Defining Inputs and Outputs

An important step in efficiency analysis is the definition of inputs and outputs and their prices. Indeed, the results can be misleading or meaningless if these quantities are poorly defined. This problem is especially acute in the service sector, where many outputs are intangible and many prices are implicit. Defining inputs also must be done with care in studies of the U.S. insurance industry, where basic data on some inputs, such as the number of hours worked and number of employees, are not available in public sources. In spite of the challenges, researchers have come up with ways to measure inputs, outputs, and prices that produce

economically meaningful efficiency scores. This section discusses the concepts and definitions of inputs, outputs, and input and output prices.

Inputs and Input Prices

Insurer inputs can be classified into three principal groups: labor, business services and materials, and capital. For some applications it also may make sense to split labor into agent labor and all other (mostly home office) labor because the two types of labor have different prices and are used in different proportions by firms in the industry (e.g., some firms use direct marketing in whole or in part, while others rely heavily on agents).[8] In addition, there are at least three types of capital that can be considered – physical capital, debt capital, and equity capital. However, it is rare for insurance efficiency studies to utilize more than four inputs.

A tabulation of insurer expenses by category is shown in Table 1. Insurance is a labor intensive industry, with agents' compensation accounting for about 40.9 (30.0) percent of total life-health (property-liability) insurer operating expenses and other personnel costs (claims adjusters and administrative labor) accounting for 27.0 (43.6) percent of total operating expenses. Physical capital expenditures (equipment and real estate costs) amount to 5.7 (6.3) percent of expenses for life-health (property-liability) insurers, and all other expenses – business services and materials – accounting for 24.2 (19.5) percent. Because physical capital expenditures are a small proportion of the total, they are sometimes lumped together with business services and materials.

Financial capital is also an important expenditure for insurers. E.g., the equity capital-to-asset ratios for life-health and property-liability insurers are 6.9 and 32.8 percent, respectively.[9] Accordingly, many insurance industry studies include equity capital as an input. The rationale for the use of equity capital is that insurers must maintain equity capital to back the promise to pay claims even if losses are higher than expected

[8]Some studies also have included agent labor in the business services category (e.g., Berger, Cummins, and Weiss, 1997).

[9]These ratios are based on 1996 data obtained from U.S. Board of Governors of the Federal Reserve System, Federal Reserve Flow of Funds Accounts.

and to satisfy regulatory requirements. The rationale for the use of debt capital is similar to that for the use of deposits as an input in banking, i.e., that insurers raise debt capital by issuing insurance and annuity policies and invest the capital as part of the intermediation function. However, debt capital is not always used as an input in banking or insurance studies because reserves for insurers and deposits for banks have some characteristics of both inputs and outputs. In addition, to the extent that the definition of insurance output and output price accounts for the time value of money, it is probably not necessary to include debt capital as an input.

Because physical measures of input quantities are not publicly available for insurers, the approach taken in most insurance efficiency studies is to impute the quantity of physical inputs by dividing the relevant insurer expense item by a corresponding price index, wage rate, or other type of deflator. E.g., the quantity of labor is equal to the total expenditures on labor, taken from the regulatory annual statement, divided by the wage rate; and the quantity of materials and business services is computed by dividing expenditures on these inputs by a price index.

The wage rate or price of administrative labor is usually measured for life insurers using U.S. Department of Labor (DOL) data on average weekly wages for Standard Industrial Classification (SIC) class 6311, life insurance companies, and for property-liability insurers using DOL data on SIC class 6331, property-liability companies. Because wages vary significantly by state, the ideal administrative wage rate would be a weighted average based on the amount of work performed in various locations. However, to do this accurately would require data on the locations and relative sizes of the insurer's processing operations, which are not publicly available.

Two approximations that are often used for administrative labor are the wage rate for the state in which the company maintains its home office and a weighted average wage rate using the proportions of premiums written by state as weights. Neither measure is completely satisfactory. Most insurers either conduct their operations from a single home office or rely on regional (not state) offices. The limited research available on the sensitivity of results to this wage variable suggests that the efficiency scores are not very sensitive to the

choice of administrative wage rate and that the efficiency rankings are affected even less.[10] Our view is that it makes more sense to use the wage rate for the state where the home office is located rather than the premium-weighted-average wage rate.

The price of agent labor is measured using U.S. Department of Labor data on average weekly wages for SIC class 6411, insurance agents. A weighted average wage variable is used, with weights equal to the proportion of an insurer's premiums written in each state. The weighted average approach is more appropriate for agent labor than for home office labor because most agency services are provided at the local level, whereas most of the other tasks performed by insurance company employees take place at the home office or in regional offices. The price deflator for the materials category is the business services deflator from the U.S. Department of Commerce, Bureau of Economic Analysis.

Because the data for all extant insurance efficiency studies comes from regulatory annual statements, the quantity of equity capital is usually defined as the insurer's statutory policyholders surplus. In property-liability insurance studies this is sometimes adjusted by an estimate of the equity in the unearned premium reserves and other statutory balance sheet categories such as non-admitted assets which are not consistent with generally accepted accounting principles (GAAP). Such adjustments are not possible for life insurers because of the complexity of the pre-paid expense calculations in life insurance. A possibility for future research would be to conduct the analysis using GAAP accounting

statements.[11]

---

[10]This is based on unpublished sensitivity tests conducted by the authors in the course of researching the effects of consolidation in the life insurance industry (Cummins, Tennyson, and Weiss, 1998). We conducted the analysis under three different administrative wage rate assumptions – the national average weekly wage rate for SIC class 6311, the wage rate for the state where the home office is located, and the state premium-weighted-average wage rate. The analysis showed that the three sets of results are virtually identical and do not lead to different economic conclusions. The results with the home office state wage rate are reported in the paper.

[11]Researchers in other industries use the GAAP data reported by Compustat. However, because insurance accounting differs significantly from accounting in other industries, the Compustat data on insurers is virtually worthless.

To measure the cost of equity capital, one approach would be to use the market value rate of return on equity (ROE). However, few insurers are publicly traded so using market ROE would greatly restrict the sample size. Consequently, book value measures usually are used. One approach is to use the average book ROE (net income divided by policyholders surplus) for the three or five years prior to the year of analysis. A problem with this approach is that it reduces the number of years for which efficiencies can be calculated by requiring at least three years prior to the start of the first year of efficiency analysis to compute average ROE. Another problem is that realized ROE can be negative, whereas the ex ante ROE must be positive. An alternative approach to ROE estimation is to estimate a regression equation with realized ROE as the dependent variable and variables such as leverage, business mix, asset mix, and other company characteristics as independent variables. However, the use of negative ROE values in the dependent variable is still not justified theoretically.

A method that avoids the theoretical problem of negative ROE values but does not provide for much variability in costs of capital among insurers is to base the cost of capital on the financial ratings assigned by the A.M. Best Company, the leading financial rating firm for insurers. For example, Cummins, Tennyson, and Weiss (1998) adopt a three-tier approach to measuring the cost of capital based on Best's ratings. Best's uses a fifteen tier letter-coded rating system ranging from A++ for the strongest insurers to F for insurers in liquidation. The three tiers consist of the four ratings in the "A" range, the four ratings in the "B" range, and all other rating categories. Based on an examination of the equity cost of capital for traded life insurers, a cost of capital of 12 percent is assigned to the top tier, 15 percent for the middle tier, and 18 percent for insurers in the lowest quality-tier. As a robustness check, Cummins, Tennyson, and Weiss (1998) also conduct their analysis using the insurers' average return on book equity over the three years prior to each sample year. Although the use of the alternative cost of capital measure did not materially affect the results, it seems clear that more work on the cost of capital issue has the potential to improve the analysis of insurer efficiency.

The debt capital of insurers consists primarily of funds borrowed from policyholders. For life insurers, debt capital includes the aggregate reserve for life policies and contracts, the aggregate reserve for accident and

health policies, the liability for premium and other deposit funds, and other reserve items. For property-liability insurers, reserves consist of the sum of loss reserves and unearned premium reserves.[12] Insurers may borrow money through their holding companies, especially if they are publicly traded, but the amount of borrowed funds appearing on the statutory annual statements is trivial in comparison with policy reserves and thus is generally lumped together with reserves.

The interest payment made to policyholders for the use of policyholder-supplied debt capital (i.e., the cost of this type of capital) is implicit in the premium and in the dividend payments made by insurers to policyholders. The cost of policyholder-supplied debt capital is estimated as the ratio of total expected investment income minus expected investment income attributed to equity capital divided by average policyholder-supplied debt capital (Berger, Cummins, and Weiss, 1997). Expected investment income attributable to equity capital equals the expected rate of investment return multiplied by average equity capital for the year. This is based on the Myers and Cohn (1987) argument that investors will not supply capital to an insurer unless they receive a market return equal to the amount they could receive by investing in an asset portfolio that replicates the insurer's portfolio plus a risk premium for any additional costs associated with committing capital to the insurance business.

Outputs and Output Prices

Measuring Financial Services Output. Insurers are analogous to other firms in the financial sector of the economy in that their outputs consist primarily of services, many of which are intangible. Three principal approaches have been used to measure outputs in the financial services sector: the asset or intermediation approach, the user-cost approach, and the value-added approach (see Berger and Humphrey, 1992b). The

[12]The unearned premium reserve is often reduced by an estimate of prepaid expenses. Under statutory accounting rules, insurers are required to maintain reserves equal to 100 percent of unearned premiums, even though they have already paid a substantial proportion of the commissions and administrative costs covered by the expense component of the premium. We use a standard GAAP accounting adjustment for prepaid expenses, adding back to equity the following amount: UPR_t*(1-.5*(loss ratio(t)+loss ratio(t-1)), where UPR_t is the unearned premium reserve at the end of year t, and loss ratio(t) is the loss and loss adjustment expense ratio for year t.

asset approach treats financial service firms as pure financial intermediaries, borrowing funds from one set of decision makers, transforming the resulting liabilities into assets, and receiving and paying out interest to cover the time value of funds used in this capacity. The asset approach would be inappropriate for property-liability insurers because they provide many services in addition to financial intermediation. In fact, the intermediation function is somewhat incidental to property-liability insurers, arising out of the contract enforcement costs that would be incurred if premiums were not paid in advance of covered loss events. This is true to a lesser extent for life insurers, where intermediation is the most important function. However, ignoring insurance outputs is likely to overlook important distinctions among insurers and thus give less accurate results than if a wider range of outputs were used. Accordingly, the asset approach also is not optimal for life insurers.

The user-cost method determines whether a financial product is an input or output on the basis of its net contribution to the revenues of the financial institution (Hancock, 1985). If the financial returns on an asset exceed the opportunity cost of funds or if the financial costs of a liability are less than the opportunity costs, then the product is considered to be a financial output. Otherwise, it is classified as a financial input. This method is theoretically sound but requires precise data on product revenues and opportunity costs, which are difficult to estimate.[13] It is particularly inaccurate in industries such as property-liability insurance, because insurance policies bundle together many services (risk pooling, claims settlement, intermediation, etc.), which are priced implicitly.

The third approach to measuring output – the value-added approach – is the most appropriate method for studying insurance efficiency. The value-added approach considers all asset and liability categories to have some output characteristics rather than distinguishing inputs from outputs in a mutually exclusive way. The categories having significant value-added, as judged using operating cost allocations, are employed as important outputs. Others are treated as unimportant outputs, intermediate products, or inputs,

[13]Efforts to apply the user cost method in banking found that the classifications of inputs and outputs were not robust to the choice of opportunity cost estimates nor were they robust over time (see Berger and Humphrey, 1992b).

depending on the characteristics of the specific activity under consideration. The following discussion will focus solely on the value-added approach.

Services Provided by Insurers. Since insurance outputs are mostly intangible, it is necessary to find suitable proxies for the quantities of services provided by insurers. This section discusses the principal services provided and subsequent sections deal with theoretical and practical aspects of insurance output measurement.

Insurers provide three principal services:

- **Risk-pooling and risk-bearing.** Insurance provides a mechanism for consumers and businesses exposed to insurable contingencies to engage in risk reduction through pooling. Insurers collect premiums from their customers and redistribute most of the funds to those policyholders who sustain losses. The actuarial, underwriting, and related expenses incurred in operating the risk pool are a major component of value added in insurance. Policyholders may also have their risks reduced because some of these risks are borne by shareholders of the insurance company (for stock companies), by previous policyholders whose capital has been left in the company (for mutual organizations), or by other parties holding the debt of the insurance company (for both groups). Again, this creates value-added by increasing economic security.

- **"Real" financial services relating to insured losses.** Insurers provide a variety of real services for policyholders. In life insurance, these services include financial planning and counseling for individuals and pension and benefit plan administration for businesses. In property-liability insurance, real services include risk surveys to identify unusual loss exposures, the design of programs to cover these and other risks, and recommendations regarding deductibles and policy limits. Insurers also provide loss prevention services such as programs to reduce the incidence of employment-related injuries. Loss settlement services include valuation of property losses, negotiations with contractors, and legal representation for liability claims. By contracting with insurers to provide these services, policyholders can take advantage of insurers' extensive experience and specialized expertise to reduce costs associated with insurable risks.

- **Intermediation.** Insurers issue debt contracts (insurance policies and annuities) and invest the funds until they are withdrawn by policyholders (in the case of asset accumulation products sold by life insurers) or are needed to pay claims. In life insurance, interest credits are made directly to policyholder accounts to reflect investment income; whereas, in property-liability insurance, policyholders receive a discount in the premiums they pay to compensate for the opportunity cost of the funds held by the insurer, analogous to interest payments on corporate debt. The borrowed funds are invested primarily in marketable securities; and the intermediation process often involves investing in asset classes such as privately placed bonds that are not available to the public. The net interest margin between the rate of return earned on assets and the rate credited to policyholders represents the value-added of the intermediation function.

Obtaining precise information on value-added in insurance is difficult because publicly available data do not break down costs according to the services provided. Nevertheless, some rough estimates are available

to help us identify outputs. In 1996, about 41 (30) percent of operating expenses for life insurers (property-liability insurers) were for agents' commissions. Agents perform real services such as financial counseling and giving advice on coverages and deductibles. They also collect underwriting information and expand the size and presumably the diversification of the insurer's risk pool, both of which are associated with the risk-pooling function. About 27 (28.9) percent of total expenses are for personnel costs for functions other than sales, claims settlement, and investments. These expenditures are for the underwriters, actuaries, and administrators that operate the insurance risk pool and thus are primarily attributable to the risk-pooling function. For property-liability insurers, a substantial share of expenses (14.7 percent) goes for claims settlement services, which include such real services as providing a legal defense against liability suits. Investment expenses account for 9 (2.1) percent of total expenses for life (property-liability) insurers. These expenses along with the net interest margin between what insurers earn on their investments and what they credit to policyholders, is a measure of the value added by the intermediation function. Although insurers do not disclose the net interest margin, a rough idea of the potential magnitude of this component of value-added can be obtained by observing that a 50 basis point margin on invested assets would be equivalent to 13.6 (10.2) percent of total expenses for life (property-liability) insurers. Thus, intermediation is also an important output for insurers.

Defining Insurance Output: Theoretical Foundations. Before turning to the specification of the variables used to represent insurer outputs in efficiency estimation, we briefly consider the concept of insurance output from a theoretical perspective. The provision of real services poses no conceptual hurdles that need to be explored here. However, it is useful to explore the concept of the value-added from the risk-pooling/risk-bearing function in the context of the theory of insurance economics. The treatment of the intermediation function also requires some discussion.

In terms of insurance economics, the value-added from risk-pooling is measured by the Pratt-Arrow concept of the *insurance premium.* The result is stated succinctly by Arrow (1971, p. 95):

> Consider an individual faced with a random outcome Y and offered the alternative of a certain income, Y_0. A risk averter would be willing to accept a value of Y_0 less than the mean value, E(Y), of the random income; the difference may be thought of as an insurance premium.

Stated more precisely, the insurance premium (value-added) is the amount which makes the individual just indifferent between retaining and insuring the risk, i.e., the insurance premium π is the solution to the equation:

$$U(W-\mu_L-\pi) = E(U(W-L)) = \int U(W-L)\, f(L)\, dL \qquad \textbf{(12)}$$

where $U(W)$ = utility function, with $U' > 0$, $U'' < 0$

W = initial wealth (non-stochastic),

L = the loss (stochastic), with $L \geq 0$,

$f(L)$ = the probability of loss distribution, and

$\mu_L = E(L)$.

Thus, the value added by the insurance transaction is the maximum amount over and above the expected loss the policyholder is willing to pay, i.e., π. After all, the consumer clearly has the option of going uninsured and having the risky expected wealth $W-\mu_L$. It is the additional amount he/she is willing to pay that constitutes the value of the insurance.

In a competitive market, the full amount of consumer welfare gain from insurance may not be observed, i.e., the market may be able to provide the insurance for a loading less than π. It is not possible to measure the unobservable consumers surplus that results. However, it should be clear that the amount paid in addition to the expected value is the measurable value added by risk-pooling.

Although we have used the term insurance premium in this discussion to be consistent with Arrow (1971), in the remainder of the paper we refer to π as the *loading* in order to avoid confusion with the standard terminology in the insurance literature, where the term *premium* is used to mean the total amount paid by the policyholder for insurance, i.e., the expected loss plus the loading.

Because premiums are usually paid in advance of loss payments, it is necessary to appropriately account for investment income when measuring insurance output, output prices, revenues and profits. The correct approach for incorporating investment income can be illustrated by a simple one-period, two-date model of the insurance firm. The insurer is assumed to commit equity capital of S to the insurance enterprise

at time 0. Premiums in the amount P are paid at time zero, and the premiums and equity are invested at rate of return r. Losses are paid at the end of the period (time 1). To avoid unnecessarily complicating the analysis, we assume that there are no taxes.[14]

The first concept to illustrate is the price of insurance, which corresponds to π in equation (12). Following the approach in Myers and Cohn (1987) and Cummins (1990), the premium is:

$$P = \frac{L(1+e) + S\rho}{1+r} \qquad \textbf{(13)}$$

where L = the expected loss,

e = insurer expenses expressed as a proportion of the expected loss, and

ρ = the risk premium received by equity holders for bearing insurance risk.

The quantity of insurer output is proxied by the present value of losses incurred, i.e., output = L/(1+r). This is appropriate because the purpose of insurance is to redistribute funds from those members of the pool who do not have a loss to those who do suffer a loss. Thus, L is the total amount redistributed by the insurer. Insurer revenues are equal to total premiums received plus investment income earned, i.e., revenues = P + r(P+S); and value-added is defined as revenues minus loss payments and the interest earned on equity, or

$$ValueAdded = P + r(P+S) - rS - L = eL + \rho S \qquad \textbf{(14)}$$

It is necessary to subtract out the investment income on equity because this amount will be earned by equity holders in any case. Equity holders have the option of writing no insurance and thus operating as a mutual fund so that merely investing the equity carries no opportunity costs associated with operating an insurance business. The additional costs resulting from placing the money at risk in the insurance business are reflected in the risk premium ρ. The total value-added, eL + ρS, thus equals the insurers expenses plus the owners' profit charge for bearing insurance risk. The price of insurance is defined as

[14]The model can be easily generalized to incorporate taxes. See Myers and Cohn (1987) and Cummins (1990).

the value-added per dollar of output, i.e.,[15]

$$Price = \frac{P - PV(L)}{PV(L)} = \frac{P - \frac{L}{1+r}}{\frac{L}{1+r}} = e + \frac{S}{L}\rho \qquad \textbf{(15)}$$

This result can easily be generalized to incorporate the intermediation function. This is done by discounting at a rate $r_P < r$ to obtain the premium, where $(1+r) = (1+r_P)(1+m)$ and m = the net interest margin received by the insurer for performing the intermediation function. Continuing to use r as the investment income rate, it is easily shown that the value-added becomes:

$$V = m[L(1+e) + \rho S] + [eL + \rho S] \qquad \textbf{(16)}$$

which equals the value added from intermediation plus the value added by risk-pooling.

Defining Insurance Output In Practice. Some efficiency studies have used premiums to measure output. This is a fallacy, however, because premiums represent price times the quantity of output not output (Yuengert, 1993). Thus, it is necessary to develop measures that are consistent with the preceding discussion.

For property-liability insurers, it is possible to develop practical measures of price and output that correspond closely to the theoretical measures discussed above. Specifically, the present value of real losses incurred can be used as a reasonable proxy for output. Estimates of the payout proportions can be obtained by applying the Internal Revenue Service or Taylor separation methods to data from Schedule P of the regulatory annual statement that provides information on reserve runoffs;[16] and discounting can be performed

[15]It is hoped that this discussion will clear up some confusion in the literature about insurance price and output. For example, Armknecht and Ginsburg (1992) define insurance price as (P-L)/L, which in our notation equals [e+(S/L)ρ-r]/(1+r), i.e., in their formulation investment income on policyholder funds is part of the price. This is not correct because the policyholder could invest these funds in any event and thus there is no opportunity cost associated having the funds invested instead by the insurer except in the case where the insurer earns a net interest margin from intermediation. The type of fallacious reasoning that the Armknecht-Ginsburg definition can lead to is exemplified by Lipsey (1992), who, in commenting on their paper observes that "a rise in investment earnings by casualty insurance companies does make it cheaper to insurer your car."

[16]For further discussion of the estimation of payout proportions, see Cummins (1990).

using U.S. Treasury yield curves. Because the various lines of business offered by insurers have different risk and payout characteristics, it is usually appropriate to use several output measures, grouping together lines with similar characteristics. Output prices can then be obtained using the formula: (P-PV(L))/PV(L) as in equation (15).

For life insurers, it is not possible to obtain meaningful present values based on publicly available data because of the complexity of life insurance products and limitations on the types of information reported by life companies. The approach used in most recent papers is to define output as incurred benefits plus additions to reserves (e.g., Yuengert, 1993, Cummins, Tennyson, and Weiss, 1998). Incurred benefits represent payments received by policyholders in the current year and are useful proxies for the risk-pooling and risk-bearing functions because they measure the amount of funds pooled by insurers and redistributed to policyholders as compensation for insured events. Most life insurance and annuity products involve the accumulation of assets, either to pay future death benefits or to be received as income through an annuity. The funds received by insurers that are not needed for benefit payments and expenses are added to policyholder reserves. Additions to reserve thus should be highly correlated with the intermediation output. Both incurred benefits and additions to reserves are correlated with real services provided by insurers, such as benefit administration in the case of group insurance and financial planning in the case of retail products. Because the major products offered by life insurers differ in the types of contingent events that are covered and in the relative importance of the risk-pooling, intermediation, and real service components of output, it is necessary to define several types of output, representing the major lines of insurance. In keeping with the value-added approach to output measurement, the price of each insurance output is defined as the sum of premiums and investment income minus output for the line divided by output.[17]

It is also possible to use physical measures to proxy for insurance outputs in life insurance. Life insurers are required to report the number of death and other types of benefits paid and incurred as well as the

[17]Life insurers are required to allocate investment income by line in their regulatory annual reports, and we use the reported allocations in defining output prices.

number of policies issued and in force and the amount of insurance written and in force. Using these physical measures has the potential to allow for differences among companies in a more detailed way and thus to provide better estimates of efficiency. For example, it is known that expenses differ among companies as a function of average policy sizes and the proportion of new business to existing business, but the usual measures of life insurance output do not control for such differences. Estimating life insurer efficiency with physical output measures thus is a promising avenue for future research.

4. A Survey of Insurance Efficiency Research

This section provides a survey of the research on productivity and efficiency in the insurance industry. We limit the analysis to studies that utilize modern frontier efficiency methodologies and attempt to provide a comprehensive survey of this literature.

Outputs and Inputs

The outputs used in the extant insurance efficiency studies are summarized in Table 2. The life insurance studies are discussed first, followed by property-liability. While some of the earlier life insurance studies used premiums as an output measure (e.g., Fecher, 1993, Gardner and Grace, 1993), most of the more recent studies have corrected this error and used more appropriate output measures. The emerging consensus in the literature is that incurred benefits and changes in reserves should be used to measure life insurance output. This measure is used by Cummins, Tennyson, and Weiss (1998), Cummins and Zi (1998), Meador, Ryan, and Schellhorn (1997), and Kim and Grace (1995). Yuengert (1993) uses additions to reserves but does not include incurred benefits. Fukuyama (1997), following an intermediation approach to defining output, uses reserves and loans as his output measures. Another group of authors uses physical output measures such as numbers of policies and/or insurance in force (Bernstein, 1997, Weiss, 1986, Kellner and Mathewson, 1983). We are aware of no research that compares the use of monetary and physical output proxies in measuring insurer output.

Nearly all extant property-liability insurance studies use either the present value of losses or losses as an output measure, usually broken down into four or more lines of insurance. Berger, Cummins and Weiss

(1997), Cummins, Weiss, and Zi (1998), and Cummins and Weiss (1993) use present values, whereas Weiss (1991a, 1991b, 1990) uses undiscounted losses. Some papers also use assets to measure the intermediation function, while others use policyholder reserves. Only one study (Fecher, et al., 1993) uses premiums as output for property-liability insurers.

There is even more uniformity in the choice of inputs for insurance efficiency studies than there is for the choice of outputs (see Table 3). Virtually every study uses labor and capital as well as a third category called business services or materials to cover inputs not included in the labor and capital categories. About one-fourth of the studies distinguish between home office and agent labor. The studies are about evenly split between the use of physical and financial capital, and two studies use both physical and financial capital. A considerable amount of agreement also exists about the types of wage and price indices that are used to represent prices of the inputs. It would be useful for future research to compare efficiency scores using alternative definitions of the inputs.

Average Efficiency Scores

The results of the insurance industry efficiency studies are summarized in Table 4. The majority of the studies focus on the U.S., but analyses also have been conducted for France, Italy, and Japan. Two of the fourteen studies summarized in Table 4 use both econometric and mathematical programming methodologies. Of the remaining twelve studies, two-thirds employ econometric techniques while the remainder use mathematical programming. The mathematical programming technique used almost exclusively has been DEA. Only one study utilized the FDH approach, although it is to be emphasized that FDH is relatively new. Given the potential importance of the convexity assumption, it would be useful to have additional research comparing the two mathematical programming approaches. The single study that made such a comparison for insurers found that the FDH efficiency scores tended to correlate somewhat better than the DEA scores with conventional performance measures such as return on equity, whereas both types of mathematical programming scores correlated somewhat better with the conventional measures than did the econometric scores (Cummins and Zi, 1998).

As is true for efficiency analyses in general, the majority of applications have estimated cost and/or technical efficiency. Eleven of the fourteen studies present cost efficiency results, and five studies present technical efficiency results. Two studies consider profit efficiency, and only one reports revenue efficiency results. As efficiency analysis evolves, the trend will be towards the estimation of both technical/cost efficiency and revenue/profit efficiency. One reason for this is that estimating only cost or technical efficiency misses the "big picture" question, i.e., whether the firm characteristics under analysis have an impact on the bottom line.

A related reason for investigating revenue/profit efficiency as well as technical/cost efficiency is that looking at the latter types of efficiency alone can produce misleading conclusions. An example is the analysis of independent and exclusive agency distribution systems conducted by Berger, Cummins, and Weiss (BCW) (1997). Prior researchers consistently found that independent agency insurers had higher expense ratios than direct writing insurers and almost invariably interpreted the results as implying that independent agents were less efficient. BCW provide evidence that this is a mistaken conclusion. Rather than being due mainly to inefficiency, the higher costs of independent agents appear to be due to unmeasured product quality differences. These differences are made up on the revenue side so that there are no significant differences in profit efficiency among insurers using the two distribution systems. Thus, estimating both cost and profit efficiency provides a general technique to control for unmeasured differences in the quality of services provided.

The average cost efficiency estimates for life insurers are reasonably consistent across studies. Of the six U.S. studies that report averages, four report average cost efficiencies between 0.35 and 0.5 and a fifth (Cummins and Zi, 1998) reports scores in this range for some methodologies. Higher scores were reported by Cummins and Zi (1998) for their econometric models and by Grace (1995), also using econometrics. DEA scores are generally expected to be lower than econometric scores, ceteris paribus, because DEA measures all departures from the frontier as inefficiency, whereas the econometric approach allows for random error. The higher scores for DFA also are expected given the relaxation of the convexity condition.

The differences among the efficiency scores produced by various methodologies in the same study are

not necessarily a matter for concern. In most cases, the researcher is interested in ranking firms in terms of performance rather than in the absolute scores themselves. Thus, the rank correlations between various methods are usually of more interest than the average efficiencies. Cummins and Zi (1998) find that the bivariate rank correlations among scores produced by alternative econometric methods are quite high, typically above 0.95. The rank correlations between the econometric and mathematical programming methods also are reasonably high, about 0.6 on average, and the correlations between DEA and FDH are in the same range. These somewhat lower correlations imply that it is advisable to use more than one type of method when estimating efficiency.

Economies of Scale and Scope

The results with respect to scale and scope economies are presented in Table 5. The table summarizes the results of studies that use modern frontier efficiency techniques, with the exception of Kellner and Mathewson (1983) and Grace and Timme (1992), which are included because they have received a significant amount of attention in the literature. The studies are in agreement that relatively small U.S. life insurers tend to be operating with increasing returns to scale (IRS). There is less agreement regarding the asset size range where returns to scale disappear and also about the issue of whether larger insurers tend to operate with constant returns to scale (CRS) or decreasing returns to scale (DRS). The most recent studies, Cummins, Tennyson and Weiss (CTW) (1998) and Cummins and Zi (1998), find that the majority of life insurers operate with IRS or CRS up to about $1 billion in real assets and that the majority of firms with more than $1 billion in assets exhibit DRS. Yuengert (1993) and Grace and Timme (1992) find that large insurers tend to exhibit CRS. Grace and Timme find that increasing returns to scale disappear for the largest quartile of stock insurers but that increasing returns are present for all quartiles of mutuals. They do not report quartile boundaries, but based on the Cummins-Tennyson-Weiss study, the largest size quartile begins at roughly $1 billion. Thus, Grace and Timme's results are consistent with CTW in terms of the region in which returns to scale disappear. Yuengert (1993), on the other hand, finds that returns to scale do not disappear until around $15 billion in assets. Based on the CTW asset size distribution, $15 billion is approximately the 98th percentile.

The CTW study focused on the effects of consolidation on efficiency and used scale economies primarily as an explanatory variable. Consequently, scale economy results were not actually presented in their paper. Because scale economies is an important topic and CTW estimated the most detailed set of results currently available, we present their scale economies results in Table 6. The Cummins-Zi (1998) results are also reproduced in the table for purposes of comparison. The table shows the percentage of firms operating with increasing, constant, and decreasing returns to scale in each of ten asset size categories. There is a very definite inverse relationship between asset size and the percentage of firms operating with IRS and a corresponding direct relationship between size and the proportion of firms operating with DRS. Interestingly, a significant proportion of the largest firms manage to operate with CRS even though the majority of firms with more than $1 billion in assets exhibit DRS. A useful topic for future research would be to investigate the characteristics of the large CRS firms relative to the large DRS firms to try to identify the "best practices" that enable the former group to avoid DRS.

The returns to scale results are important because of the consolidation that is taking place in the life insurance industry. If scale economies disappear at about $1 billion, then it is difficult to justify mergers in the top quartile of the life insurance industry, at least on the grounds of cost economies. And if firms with more than $1 billion tend to encounter decreasing rather than constant returns to scale, mergers and acquisitions in this size range become even more difficult to justify. More research is clearly needed to resolve this issue; and more research is needed on scale economies in the property-liability insurance industry, which is also undergoing significant consolidation.

The issue of scope economies is also important because of the increasing prevalence of cross-industry mergers involving life insurers, property-liability insurers, and other financial institutions. Unfortunately, the existing research on scope economies is quite limited and not informative with respect to cross-industry mergers. Neither Yuengert (1993) nor Grace and Timme (1992) find evidence of scope economies in the U.S. life insurance industry. Kellner and Mathewson (1983) find some evidence of scope economies in the earlier years of their sample period but not for the most recent year. Their study is dated, however, ending in 1976.

Although not strictly speaking a scope economies study, Meador, Ryan, and Schellhorn (1998) analyze the efficiency of firms that are well-diversified across multiple product lines in comparison with those that follow a more focused strategy and find that diversification is associated with higher cost efficiency. This finding is intriguing, but in general it is clear that scope economies is significantly under-researched.

Total Factor Productivity Growth

Several papers explore the issue of total factor productivity (TFP) growth in the insurance industry. Measuring TFP growth is important to gauge the effects of changing industry structure such as the wave of mergers and acquisitions currently underway in the U.S. insurance industry. It is also important to measure the effects of changes in management practices and the introduction of new technologies. The two principal approaches to measuring TFP growth are the Malmquist index method and the calculation of TFP indices.

The results of TFP studies in the insurance industry are summarized in Table 7. Four papers have utilized the Malmquist approach, three have used the index approach, and one uses an econometric approach. Cummins, Tennyson, and Weiss (1998) find productivity growth of 4.1 percent per year in the U.S. life insurance industry for the period 1991-1994, while Cummins, Weiss, and Zi (1998) find virtually no growth in productivity in the U.S. property-liability insurance industry for the period 1981-1990. It is possible that advances in technology have led to higher productivity gains for the property-liability insurance industry during the 1990s, and this would be an interesting topic for future research.

Malmquist index analyses of Japanese life insurers (Fukuyama, 1997) and Italian life and property-liability insurers (Cummins, Turchetti, and Weiss, 1996) show efficiency gains that are considerably higher than in the U.S. Fukuyama reports TFP gains of about 19 percent for Japanese life insurers over the period 1988-1993. Cummins, Turchetti, and Weiss find that firms which were in the Italian market for the entire period 1986-1993 showed TFP gains of about 3.4 percent per year but that when firms that entered or exited are included in the sample, efficiency gains are about 19 percent per year. The index methodology studies tend to show more modest efficiency gains for U.S., Canadian, Japanese, and three European countries.

Other Economic Hypotheses

Efficiency analysis has been used to investigate a number of economic hypotheses and issues in addition to economies of scale and scope and total factor productivity. Several of the more important studies are summarized in Table 8. One important issue that has been investigated is the effect of organizational form on performance. The two major hypotheses about organizational form are the *expense preference* hypothesis (Mester, 1991) and the *managerial discretion* hypothesis (Mayers and Smith, 1988). Both hypotheses derive from agency theory, and they are not mutually exclusive. Both hypotheses are based on the observation that the mutual organizational form provides weaker mechanisms for owners to control managers than the stock organizational form. The expense preference hypothesis holds that mutuals will be less efficient than stocks because managers will behave opportunistically, engaging in higher perquisite consumption than stock managers because of weaker incentives to align their interests with those of owners.

The managerial discretion hypothesis posits that mutuals will be more successful in lines of business and other dimensions of firm activity that involve relatively low managerial discretion such as lines with standardized policies and good actuarial tables. The hypothesis predicts that stocks are more likely to succeed in lines where managers need more discretion such as complex commercial coverages and international operations. The managerial discretion hypothesis implies that stocks and mutuals will use different technologies, where technology is defined as including all of the contractual relationships that comprise the firm, as well as physical technology choices. Stocks and mutuals will be sorted into market segments where they are relatively successful in dealing with various types of agency costs, and efficiency differences between the two groups of firms are not necessarily predicted.

The most comprehensive study of the efficiency of U.S. mutual and stock insurers is Cummins, Weiss, and Zi (CWZ) (1998), who analyze the property-liability industry. They estimate a pooled frontier, including both organizational forms, and separate frontiers for mutuals and stocks. Hypothesis tests show that stocks and mutuals are using different technologies, supporting the managerial discretion hypothesis. They also innovate by performing *cross-frontier* analysis, i.e., computing efficiencies of mutuals (stocks) against a reference set

consisting of all stock (mutual) firms. If the distance of mutuals (stocks) from the stock (mutual) frontier is greater than the distance from their own frontier, the implication is that the stock technology dominates the mutual technology.

CWZ find that the stock technology dominates the mutual technology for producing stock output vectors, and that the mutual technology dominates the stock technology for producing mutual output vectors. This supports the managerial discretion hypothesis prediction that firms are sorted into market segments where they have comparative advantages. However, when performing the cross-frontier *cost* efficiency analysis, they find that the stock cost frontier dominates the mutual cost frontier, implying that mutuals are less successful in minimizing costs. Thus, the paper provides support for the managerial discretion hypothesis in terms of technology, but also supports the expense preference hypothesis. This type of efficiency analysis thus enables researchers to come to a much richer understanding of differences in firm performance than the conventional approach of using a single dummy variable to differentiate between stocks and mutuals.

Another sophisticated analysis of organizational form is provided by Fukuyama's (1997) study of Japanese life insurers. He finds that the Japanese stock and mutual insurers have identical technologies. Neither organizational form is clearly dominant in terms of efficiency, but the organizational forms tend to perform differently relative to one another depending on the economic conditions.

The effects of organizational form on the efficiency of U.S. life insurers have been analyzed by Gardner and Grace (1993) and Cummins and Zi (1998). Neither study finds significant efficiency differences between stocks and mutuals. However, it would be interesting to conduct a cross-frontier analysis to determine if a more sophisticated approach would find differences in efficiency by organizational form. Fecher, et al. (1993) present summary statistics by organizational form for French life and non-life insurers. Although they do not test for statistical significance, stock life insurers have higher average efficiency scores than mutuals, but mutual property-liability insurers have higher efficiencies than stocks.

The effect of insurance distribution systems on efficiency has been studied by Cummins, Turchetti, and Weiss (CTcW) (1996) and Berger, Cummins, and Weiss (BCW) (1998). CTcW find significant efficiency

differences between direct writing and independent agency firms in their sample of Italian insurers. However, BCW are able to shed considerably more light on this time-honored area of empirical investigation, as discussed above, finding that the higher costs of U.S. independent agency firms are due to unmeasured differences in service intensity that are compensated for by higher revenues.

Another application of efficiency analysis in insurance is to analyze the efficiency effects of mergers and acquisitions (M&A) (Cummins, Tennyson, and Weiss, 1998). CTW analyze the efficiency effects of mergers and acquisitions in the U.S. life insurance industry, covering the period 1988-1995. They find that acquisition targets tend to show significantly larger efficiency gains between the post and pre-acquisition periods than a control group of firms that have not been involved in M&A activity. They further find that acquiring firms prefer to acquire target firms that are operating with non-decreasing returns to scale. Finally, Weiss (1990) and Carr (1997) use productivity and efficiency estimates to investigate the effects on firm performance of regulation and management strategies, respectively. These studies provide additional examples of the important economic issues that can be investigated using efficiency analysis.

5. Summary and Conclusions

Modern frontier efficiency methodologies are rapidly becoming the dominant approach to measuring firm performance. These methodologies estimate efficient technical, cost, revenue, and profit frontiers by comparing each firm in the industry to a reference set consisting of all other firms. The frontiers can thus be considered "best practice" frontiers. Frontier efficiency methods have been applied to analyze a wide range of industries and public entities in many different nations.

The two primary methods for estimating efficient frontiers are the econometric approach and the mathematical programming approach. Both methods continue to be used extensively in the efficiency literature. Because each has advantages and disadvantages, it is advisable to estimate efficiency using more than one method. The econometric approach involves estimating a cost, revenue, or profit function, while the mathematical programming approach is usually implemented using linear programming. The mathematical programming approach provides a particularly convenient method for decomposing cost efficiency into pure

technical, scale, and allocative efficiency.

There are many important applications of frontier efficiency methods. One important application is the measurement of scale and scope economies. Measuring scale and scope economies is particularly important when industry structure is changing rapidly due to mergers, acquisitions, solvency problems, and other factors. Another important application is to measure the growth in total factor productivity (TFP). TFP growth can then be analyzed for correlations with various macro and micro-economic conditions and characteristics to determine the drivers of productivity in an industry or economy. Frontier efficiency analysis also is important in testing hypotheses about firm and industry structure, such as the effects of organizational form, product distribution systems, and corporate governance. This type of analysis often leads to a much richer understanding of the topic under investigation than more conventional approaches. Another use of efficiency analysis is in comparing performance of departments, divisions, or profit centers within a firm. Mathematical programming is particularly useful for this purpose because it is not as demanding in terms of degrees of freedom as the econometric approach. Regulation is another important area that potentially benefit from efficiency analysis. The Federal Reserve has used efficiency analysis to study the effects of bank branching, mega-mergers, and other elements of banking industry structure. This type of analysis could be used in insurance to study industry consolidation, expense and rate regulation, and solvency regulation. Efficiency and productivity analysis also has been used in cross-national comparisons of efficiency of firms and other institutions.

An important trend in the literature is to estimate profit and/or revenue efficiency in addition to technical and cost efficiency. Technical and cost efficiency are useful in studying the efficiency effects of firm characteristics and of new policies, strategies, and technologies. However, the ultimate test of any organizational feature is its impact on the bottom line, i.e., ultimately on profit. It is clearly possible to introduce a new strategy or technique in one area of the firm that improves cost efficiency which never finds its way to the bottom line due to inefficiencies in other sectors of the firm. The only way to tell whether a program has met with ultimate success is to measure its effects on profit efficiency.

A wide range of under-researched insurance topics provide fruitful avenues for future research. Economies of scale in the property-liability insurance industry has received almost no attention using modern frontier efficiency methods. The issue of organizational form in the life insurance industry could be addressed using the sophisticated cross-frontier approach to provide an in-depth test of the expense preference and managerial discretion hypotheses. Analyzing the efficiency of life insurance distribution systems using cost and profit functions could determine whether unmeasured product quality differences exist in the life insurance industry. The effects of consolidation on efficiency in the property-liability insurance industry also would be an interesting topic with both academic and regulatory implications. A final example of potential future research would be an analysis of board composition and other aspects of corporate governance on efficiency in the insurance industry.

References

Aly, H. Y., R. Grabowski, C. Pasurka, and N. Rangan, 1990, "Technical, Scale, and Allocative Efficiencies in U. S. Banking: An Empirical Investigation," *Review of Economics and Statistics* 72: 211-218.

A.M. Best Company, 1994, *Best's Aggregates and Averages: 1994 Edition* (Oldwick, NJ).

Armknecht, Paul A. and Daniel H. Ginsburg, 1992, "Improvements in Measuring Price Changes In Consumer Services," in Z. Griliches, ed., *Output Measurement in the Service Sectors*, National Bureau of Economic Research, Studies in Income and Wealth, Vol. 56, University of Chicago Press (Chicago, IL): 109-156.

Arrow, Kenneth, 1971, *Essays in the Theory of Risk Bearing* (Chicago: Markham Publishing Company).

Bauer, Paul W., Allen N. Berger, Gary D. Ferrier, and David B. Humphrey, 1998, "Consistency Conditions for Regulatory Analysis of Financial Institutions: A Comparison of Frontier Efficiency Methods," *Journal of Economics and Business* 50: 85-114.

Berger, Allen N, 1993, "'Distribution-Free' Estimates of Efficiency of in the U.S. Banking Industry and Tests of the Standard Distributional Assumptions," *Journal of Productivity Analysis* 4: 261-292.

Berger, Allen N., J. David Cummins, and Mary A. Weiss, 1997, "The Coexistence of Multiple Distribution Systems for Financial Services: The Case of Property-Liability Insurance," *Journal of Business* 70: 515-546.

Berger, Allen N., Diana Hancock, and David B. Humphrey, 1993, "Bank Efficiency Derived from the Profit Function," *Journal of Banking and Finance*, 17 (April): 317-47.

Berger, Allen N., and David B. Humphrey, 1991, "The Dominance of Inefficiencies over Scale and Product Mix Economies in Banking," *Journal of Monetary Economics* 28 (August): 117-48.

Berger, Allen N., and David B. Humphrey, 1992a, "Megamergers in Banking and the Use of Cost Efficiency as an Antitrust Defense," *Antitrust Bulletin* 37 (Fall): 541-600.

Berger, Allen N., and David B. Humphrey, 1992b, "Measurement and Efficiency Issues in Commercial Banking," in Z. Griliches, ed., *Output Measurement in the Service Sectors*, National Bureau of Economic Research, Studies in Income and Wealth, Vol. 56, University of Chicago Press (Chicago, IL): 245-79.

Berger, Allen N. and David B. Humphrey, 1997, "Efficiency of Financial Institutions: International Survey and Directions for Future Research," *European Journal of Operational Research* 98: 175-212.

Bernstein, Jeffrey I. 1997. "Total Factor Productivity Growth in the Canadian Life Insurance Industry: 1979-1989," CSLS Conference on Service Centre Productivity and the Productivity Paradox, April 11-12, Ottawa, Canada.

Carr, Roderick M. 1997. "Strategic Choices, Firm Efficiency and Competitiveness in the US Life Insurance Industry," Doctoral Dissertation, Wharton School, University of Pennsylvania, Philadelphia.

Caves, D., L. Christensen, and W.E. Diewert (1982), "The Economic Theory of Index Numbers and the Measurement of Input, Output, and Productivity," *Econometrica* 50 (6): 1393-1414.

Caves, Douglas W., Laurits R. Christensen, and Michael W. Tretheway, 1980, "Flexible Cost Functions for

Multiproduct Firms," *Review of Economics and Statistics* 62: 477-481.

Charnes, Abraham, William Cooper, Arie Y. Lewin, and Lawrence M. Seiford, 1994, *Data Envelopment Analysis: Theory, Methodology, and Applications* (Norwell, MA: Kluwer Academic Publishers).

Christensen, Laurits R; Jorgenson, Dale W; and Lau, Lawrence J., 1971, "Conjugate Duality and the Transcendental Production Function," *Econometrica* 255:256.

Christensen, Laurits R; Jorgenson, Dale W; and Lau, Lawrence J. Transcendental Logarithmic Production Frontiers.. *Review of Economics and Statistics*. Vol. 55 (1). p 28-45. Feb. 1973.

Cummins, J. David, 1990,"Multi-Period Discounted Cash Flow Ratemaking Models in Property-Liability Insurance," *Journal of Risk and Insurance* 57: 79-109.

Cummins, J. David and Mary A. Weiss, 1993, "Measuring Cost Efficiency in the Property-Liability Insurance Industry," *Journal of Banking and Finance* 17: 463-481.

Cummins, J. David and Mary A. Weiss, 1991, "The Structure, Conduct, and Regulation of the Property-Liability Insurance Industry," in Richard E. Randall and Richard W. Kopcke, eds., *The Financial Condition and Regulation of Insurance Companies* (Boston: Federal Reserve Bank of Boston).

Cummins, D., G. Turchetti, and M. Weiss, 1997, "Productivity and Technical Efficiency in the Italian Insurance Industry," Working paper, Wharton Financial Institutions Center, University of Pennsylvania, Philadelphia.

Cummins, D., M. Weiss, and H. Zi, 1998, "Organizational Form and Efficiency: An Analysis of Stock and Mutual Property-Liability Insurers," Working paper, Wharton Financial Institutions Center, University of Pennsylvania, Philadelphia.

Cummins, D. and H. Zi, 1998, "Measuring Economic Efficiency of the US Life Insurance Industry: Econometric and Mathematical Programming Techniques," forthcoming in *Journal of Productivity Analysis*.

Cummins, D., S. Tennyson and M. Weiss 1998. "Efficiency, Scale Economies, and Consolidation in the U.S. Life Insurance Industry," forthcoming in *Journal of Banking and Finance*.

Deprins, E., L. Simar, and H. Tulkens, 1984, "Measuring Labor Efficiency in Post Offices," in M. Marchand, P. Pestieau, and H. Tulkens, eds., *The Performance of Public Enterprises: Concepts and Measurement* (Amsterdam, North Holland).

Diewert, W. Erwin, 1995, "Functional Form Problems in Modeling Insurance and Gambling," *Geneva Papers on Risk and Insurance Theory* 20: 135-150.

Diewert, W. Erwin, 1981, "The Theory of Total Factor Productivity Measurement in Regulated Industries," in *Productivity Measurement in Regulated Industries*. T. G. Cowing and R. Stevenson (eds.) New York: Academic Press.

Färe, R., S. Grosskopf, M. Norris, and Z. Zhang, 1994, "Productivity Growth, Technical Progress, and Efficiency Change in Industrialized Countries," *American Economic Review* 1994: 66-83.

Färe, R., 1988, *Fundamentals of Production Theory* (New York: Springer-Verlag).

Färe, R., S. Grosskopf, and C. A. K. Lovell, 1985, *The Measurement of Efficiency of Production*. (Boston: Kluwer-Nijhoff).

Farrell, M.J., 1957, "The Measurement of Productive Efficiency," *Journal of the Royal Statistical Society* A 120: 253-281.

Fecher, F., D. Kessler, S. Perelman, and P. Pestieau, 1993, "Productive Performance of the French Insurance Industry," *Journal of Productivity Analysis* 4: 77-93.

Ferrier, G.D., and C.A.K. Lovell, 1990, "Measuring Cost Efficiency in Banking: Econometric and Linear Programming Evidence," *Journal of Econometrics* 46: 229-245.

Fields, Joseph A. and Neil B. Murphy, 1989, "An Analysis of Efficiency in the Delivery of Financial Services: The Case of Life Insurance Agencies, *Journal of Financial Services Research* 2: 343-356.

Fried, H. O., C. A. K. Lovell, and S.S. Schmidt, eds., 1993, *The Measurement of Productive Efficiency* (New York: Oxford University Press).

Fukuyama, Hirofumi, 1997, "Investigating Productive Efficiency and Productivity Changes of Japanese Life Insurance Companies," *Pacific-Basin Finance Journal* 122:

Gallant, A.R., 1982, "Unbiased Determination of Production Technologies," *Journal of Econometrics* 20: 285-323.

Gardner, L., and M. Grace, 1993, "X-Efficiency in the U.S. Life Insurance Industry," *Journal of Banking and Finance* 17: 497-510.

Gardner, L., and M. Grace, 1997, "Cost Differences Between Mutual and Stock Life Insurance Companies," Working paper, Center for Risk Management and Insurance Research, College of Business Administration, Georgia State University, Atlanta, GA.

Grace, Martin F., 1995, "Firm Efficiency and Insolvency: An Investigation of the US Insurance Industry," Working paper, Center for Risk Management and Insurance Research, College of Business Administration, Georgia State University, Atlanta, GA.

Grace, Martin F. and Richard D. Phillips, 1997, "The Allocation of Government Regulatory Authority within a Federal System of Government: Fiscal Federalism and the Case of Insurance Regulation," Working paper, Center for Risk Management and Insurance Research, College of Business Administration, Georgia State University, Atlanta, GA.

Grace, Martin F. and Stephen G. Timme, 1992, "An Examination of Cost Economies in the United States Life Insurance Industry," *Journal of Risk and Insurance* 59: 72-103.

Grosskopf, Shawna, 1993, "Efficiency and Productivity," in H.O. Fried, C.A.K. Lovell, and S.S. Schmidt, eds., *The Measurement of Productive Efficiency* (New York: Oxford University Press).

Grosskopf, Shawna, 1996, "Statistical Inference and Nonparametric Efficiency: A Selective Survey," *Journal of Productivity Analysis* 7: 161-176.

Grifell-Tatjé, E. and C.A.K. Lovell, 1993, "Deregulation and Productivity Decline: The Case of Spanish Savings Banks," Working Paper, University of Georgia, Athens, GA.

Halpern, P. J. and G. F. Mathewson, April 1975, "Economies of Scale in Financial Institutions: A General Model Applied to Insurance," *Journal of Monetary Economics* 1 (2): 203-220.

Hancock, Diana, 1985, "The Financial Firm: Production With Monetary and Nonmonetary Goods, *Journal of Political Economy* 93: 859-880.

Hermalin, Benjamin E. and Nancy E. Wallace, 1994, "The Determinants of Efficiency and Solvency in Savings and Loans," *RAND Journal of Economics* 25: 361-381.

Hornstein, A. and E. C. Prescott, 1991, "Measures of the Insurance Sector Output," *Geneva Papers on Risk and Insurance* 16: 191-206.

Kellner, S., and F. G. Mathewson, 1983, "Entry, Size Distribution, Scale, and Scope Economies in the Life Insurance Industry," *Journal of Business* 56: 25-44.

Kim, Hunsoo and Martin F. Grace, 1995, "Potential Ex Post Efficiency Gains of Insurance Company Mergers," Working paper, Center for Risk Management and Insurance Research, College of Business Administration, Georgia State University, Atlanta, GA.

Kwon, Wook Jean and Martin F. Grace, 1996, "Examination of Cross Subsidies in the Workers' Compensation Market," Working paper, Center for Risk Management and Insurance Research, College of Business Administration, Georgia State University, Atlanta, GA.

Land, Kenneth C., C.A. Knox Lovell, and Sten Thore, 1993, "Chance-Constrained Data Envelopment Analysis," *Managerial and Decision Economics* 14: 541-554.

Lipsey, Robert E., 1992, "Comment on Armknecht and Ginsburg," in Z. Griliches, ed., *Output Measurement in the Service Sectors*, National Bureau of Economic Research, Studies in Income and Wealth, Vol. 56, University of Chicago Press (Chicago, IL): 156-157.

Lovell, C.A.K., 1993, "Production Frontiers and Productive Efficiency," in H.O. Fried, C.A.K. Lovell, and S.S. Schmidt, eds., *The Measurement of Productive Efficiency* (New York: Oxford University Press).

Mayers, David and Clifford W. Smith, Jr., 1988, "Ownership Structure Across Lines of Property-Casualty Insurance," *Journal of Law and Economics* 31: 351-378.

McAllister, P.H. and D. McManus, 1993, "Resolving the Scale Efficiency Puzzle in Banking," *Journal of Banking and Finance* 17: 389-406.

McNamara, M. and S. R. Ghee, 1992, "The Effects of Demutualization on Firm Efficiency," *Journal of Risk and Insurance* 59: 67-79.

Meador, Joseph W., Harley E. Ryan, Jr., and Carolin d. Schellhorn, 1998, "Product Focus Versus Diversification: Estimates of X-Efficiency for the U.S. Life Insurance Industry," working paper, Northeastern University, Boston, MA.

Mester, L.J., 1991,"Agency Costs Among Savings and Loans,"*Journal of Financial Intermediation* 1: 257-278.

Mitchell, Karlyn and Nur M. Onvural, 1992, "Economies of Scale and Scope at Large Commercial Banks: Evidence from the Fourier Flexible Functional Form," *Journal of Money, Credit & Banking* 28: 178-199.

Morrison, C. J. and E.R. Berndt, 1981, "Short-Run Labor Productivity in a Dynamic Model," *Journal of Econometrics* 16: 339-65.

Myers, Stewart and Richard Cohn, 1987, "Insurance Rate Regulation and the Capital Asset Pricing Model." In J. D. Cummins and S. E. Harrington, eds., *Fair Rate of Return In Property-Liability Insurance* (Norwell, MA: Kluwer Academic Publishers).

O'Brien, C. D., 1991, "Measuring the Output of Life Assurance Companies," *Geneva Papers on Risk and Insurance* 16 (April): 207-235.

Pastor, José M., Francisco Pérez, and Javier Quesada, 1997, "Efficiency Analysis in Banking Firms: An International Comparison," *European Journal of Operational Research* 98: 395-407.

Pulley, L.B. and Y. Braunstein, 1992, "A Composite Cost Function for Multiproduct Firms With an Application To Economies of Scope in Banking," *Review of Economics and Statistics* 74: 221-230.

Pulley, L.B. and D.B. Humphrey, 1993, "The Role of Fixed Costs and Cost Complementarities in Determining Scope Economies and the Cost of Narrow Banking Proposals," *Journal of Business* 66: 437-462.

Reece, William S., 1992, "Output Price Indexes for the US Life Insurance Industry," *Journal of Risk and Insurance* 59 (March): 104-115.

Röller, Holy, 1990, "Proper Quadratic Cost Functions With an Application To the Bell System," *Review of Economics and Statistics* 72: 202-210.

Schmidt, P., and R.C. Sickles, 1984, "Production Frontiers and Panel Data," *Journal of Business and Economic Statistics* 2: 299-326.

Shephard, R. W., 1970, *Theory of Cost and Production Functions* (Princeton, N.J.: Princeton University Press).

Sherwood, Mark, 1997, "Output of the Property and Casualty Insurance Industry," Paper presented at the CSLS Conference on Service Centre Productivity and the Productivity Paradox, April 11-12, Ottawa, Canada.

Suret, J. M., 1991, "Scale and Scope Economies in the Canadian Property and Casualty Insurance Industry," *Geneva Papers on Risk and Insurance* 59: 236-256.

Vanden Eeckaut, Philippe, Henry Tulkens, and Marie-Astrid Jamar, 1993, "Cost Efficiency in Belgian Municipalities," in H.O. Fried, C.A.K. Lovell, and S.S. Schmidt, eds., *The Measurement of Productive Efficiency* (New York: Oxford University Press).

Williamson, O., 1963, "Managerial Discretion and Business Behavior," *American Economic Review* 53:1032-1057.

Weiss, Mary A., 1986, "Analysis of Productivity at the Firm level: An Application to Life Insurers," *Journal*

of Risk and Insurance (March):49-83.

Weiss, Mary A., 1987, "Macroeconomic Insurance Output Estimation," *Journal of Risk and Insurance* (September): 582-593.

Weiss, Mary A., 1989, "Analysis of Productivity at the Firm Level: An Application to Life Insurers: Author's Reply," *Journal of Risk and Insurance* (June): 341-346.

Weiss, Mary A., 1990, "Productivity Growth and Regulation of P/L Insurance: 1980-1984," *Journal of Productivity Analysis* 2 (December): 15-38.

Weiss, Mary A., 1991a, "Efficiency in the Property-Liability Insurance Industry," *Journal of Risk and Insurance* 58: 452-479.

Weiss, Mary A., 1991b, "International P/L Insurance Output, Input and Productivity Comparisons," *Geneva Papers on Risk and Insurance Theory* 16: 179-200.

Wolff E., 1991, "Productivity Growth, Capital Intensity, and Skill Levels in the US Insurance Industry, 1984-86: A Preliminary Look." *Geneva Papers*.

Yuengert, A., 1993, "The Measurement of Efficiency in Life Insurance: Estimates of a Mixed Normal-Gamma Error Model," *Journal of Banking and Finance* 17: 483-96.

Zi, Hongmin, 1994, "Measuring Economic Efficiency of US Life Insurance Industry: Econometric and Mathematical Programming Approaches," Doctoral Dissertation, Wharton School, University of Pennsylvania.

Figure 1
Production Frontier for the
Single Input-Single Output Firm

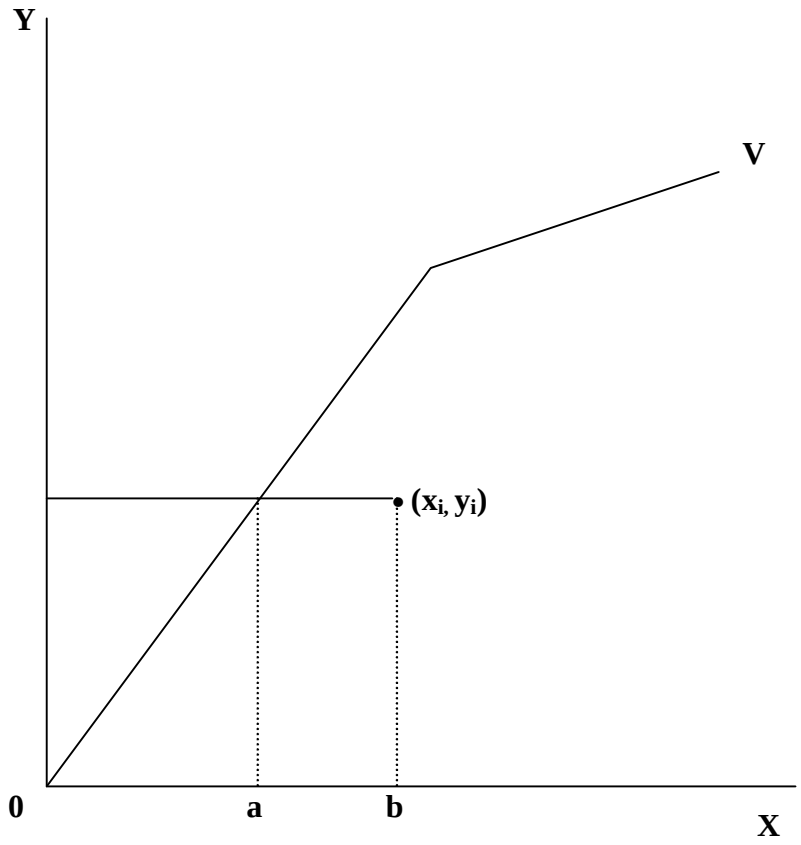


Figure 2
Farrell Technical and Allocative Efficiency

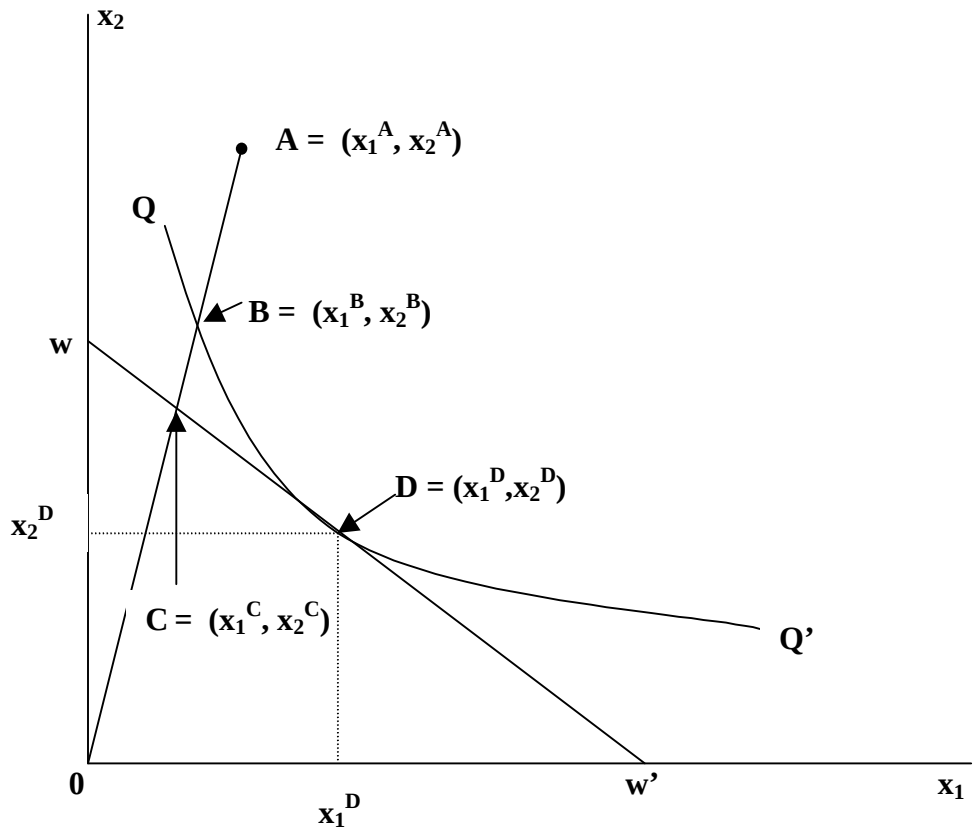


Figure 3: Pure Technical and Scale Efficiency

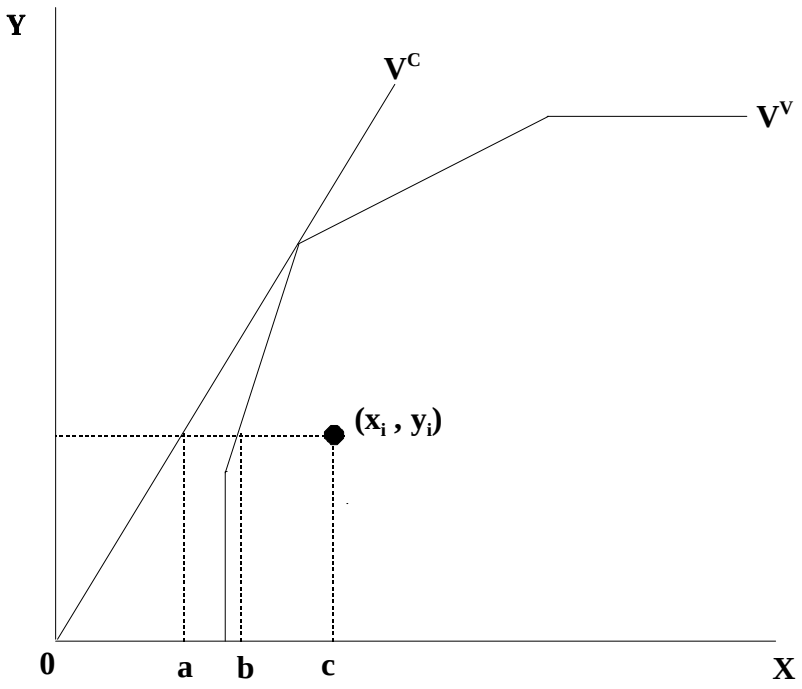
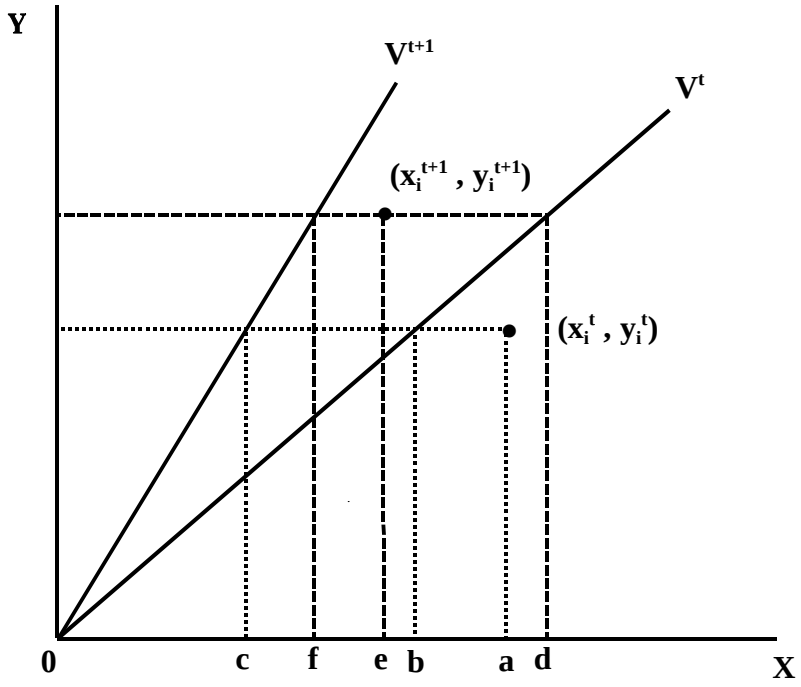


Figure 4: Productivity and Efficiency Change



$$\textit{Technical Efficiency Change} = \frac{D^t(x^t, y^t)}{D^{t+1}(x^{t+1}, y^{t+1})} = \frac{0a/0b}{0e/0f} .$$

$$\textit{Technical Change} = \left[\left(\frac{D^{t+1}(x^{t+1}, y^{t+1})}{D^t(x^{t+1}, y^{t+1})}\right)\left(\frac{D^{t+1}(x^t, y^t)}{D^t(x^t, y^t)}\right)\right]^{\frac{1}{2}}$$

$$= \left[\left(\frac{0e/0f}{0e/0d}\right)\left(\frac{0a/0c}{0a/0b}\right)\right]^{\frac{1}{2}}$$